IBM SPSS Data Preparation 24

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 반드시 35 페이지의 『주의사항』에 나오는 일반 정보를 읽으십시오.

제품 정보

이 개정판은 새 개정판에서 별도로 명시하지 않는 한, IBM® SPSS® Statistics의 버전 24, 릴리스 0. 수정사항 0 및 모든 후속 릴리스와 수정에 적용됩니다.

목차

제 1 장 데이터 준비에 대한 소개 1
데이터 준비 프로시저 사용. 1

제 2 장 검증 규칙 3
미리 정의된 검증 규칙 불러오기 3
검증 규칙 정의 3
단일-변수 규칙 정의 3
교차-변수 규칙 정의 4

제 3 장 데이터 검증. 7
데이터 기본 확인 검증 8
데이터 검증 단일-변수 규칙 9
검증 데이터 교차-변수 규칙 9
데이터 검증 결과 9
데이터 검증 저장 10

제 4 장 자동 데이터 준비 11
자동 데이터 준비 확보. 12
대화형 데이터 준비 확보 12
필드 탭. 12
설정 탭. 13
날짜 및 시간 준비 13
필드 제외 14
측정 수정 14
데이터 품질 향상 14
필드 조정 15
필드 변환 15
선택 및 구성 16
필드 이름 16
변환 적용 및 저장 17
분석 탭. 17
필드 처리 요약 18
필드. 19
동작 요약 20
예측력 20
필드 표. 21
필드 세부사항 21
동작 세부사항 22
역변환 점수 25

제 5 장 특수 케이스 식별 27
특이 케이스 결과 식별. 28
특이 케이스 식별 저장. 29
특이 케이스 결측값 식별 29
특이 케이스 식별 옵션. 29
DETECTANOMALY 명령 추가 기능 30

제 6 장 최적화 구간화 31
최적 구간화 결과 32
최적 구간화 저장 32
최적 구간화 결측값. 32
최적 구간화 옵션 32
OPTIMAL BINNING 명령 추가 기능 33

주의사항 35
상표. 37

색인. 39

제 1 장 데이터 준비에 대한 소개

컴퓨팅 시스템이 강력해짐에 따라 정보에 대한 요구도 비례하여 늘어났으며 이로 인해 점점 더 많은 데이터 콜렉션(더 많은 케이스, 더 많은 변수, 더 많은 데이터 입력 오류)이 생성되었습니다. 이러한 오류는 데이터 웨어하우징의 최종 목적인 예측 모형 예측값의 문제가 되므로 데이터를 "깨끗한 상태"로 유지해야 합니다. 그러나 데이터 검증에 대한 자동 프로세스를 구현하기 위해 필수적인 케이스 수동 확인 능력을 벗어날 정도로 데이터 웨어하우징 양이 증가되었습니다.

데이터 준비 추가 기능 모듈을 사용하면 활성 데이터 세트에 있는 특이 케이스와 유효하지 않은 케이스, 변수 및 데이터 값과 모형화를 위한 준비 데이터를 식별할 수 있습니다.

데이터 준비 프로시저 사용

데이터 준비 프로시저 사용은 특정 요구에 따라 다릅니다. 데이터를 로드한 후의 일반적인 경로는 다음과 같습니다.

- **메타데이터 준비.** 사용자 데이터 파일에서 변수를 검토하고 유효한 값, 레이블 및 측정 수준을 판별합니다. 불가능하지만 공통적으로 잘못 코딩되는 변수값 조합을 식별합니다. 이 정보를 기반으로 하여 검증 규칙을 정의합니다. 이는 시간이 걸리는 작업일 수 있지만 보통의 기반에서 유사한 속성이 있는 데이터 파일을 검증해야 하는 경우 노력할 가치가 있는 작업입니다.
- **데이터 검증.** 기본 확인 및 정의된 검증 규칙에 대한 확인을 실행하여 유효하지 않은 케이스, 변수 및 데이터 값을 식별합니다. 유효하지 않은 데이터가 있는 경우 원인을 조사하여 정정하십시오. 이를 수행하는 데 메타데이터 준비를 통한 다른 단계가 필요할 수 있습니다.
- **모형 준비.** 자동 데이터 준비를 사용하여 모형설정을 향상시키는 원래 필드의 변환을 확보합니다. 많은 예측 모형에 대한 문제를 일으킬 수 있는 잠재적인 통계 이상값을 식별합니다. 일부 이상값은 식별되지 않은 유효하지 않은 변수 값으로 인한 결과입니다. 이를 수행하는 데 메타데이터 준비를 통한 다른 단계가 필요할 수 있습니다.

데이터 파일이 "깨끗"하면 다른 추가 기능 모듈에서 모형을 작성할 준비가 됩니다.

제 2 장 검증 규칙

규칙은 케이스가 유효한지 판별하는 데 사용됩니다. 두 가지 유형의 검증 규칙이 있습니다.

- **단일-변수 규칙.** 단일-변수 규칙은 단일 변수에 적용되는 확인(범위를 벗어난 값에 대한 확인)의 고정된 세트로 구성됩니다. 단일-변수 규칙의 경우 유효한 값은 값 범위 또는 허용 가능한 값 목록으로 표현될 수 있습니다.
- **교차-변수 규칙.** 교차-변수 규칙은 단일 변수 또는 변수 조합에 적용할 수 있는 사용자 정의 규칙입니다. 교차-변수 규칙은 유효하지 않은 값을 플래그 지정한 논리식으로 정의됩니다.

검증 규칙은 사용자 데이터 파일의 데이터 사전에 저장됩니다. 이를 사용하여 규칙을 한 번 지정하면 규칙을 재사용할 수 있습니다.

미리 정의된 검증 규칙 불러오기

설치 시 포함된 외부 데이터 파일에서 미리 정의된 규칙을 불러와서 사용 준비된 검증 규칙 세트를 신속하게 얻을 수 있습니다.

미리 정의된 검증 규칙을 불러오려면 다음을 수행하십시오.

1. 메뉴에서 다음을 선택합니다.

   데이터 > 검증 > 사전 정의 규칙 불러오기...

또는 데이터 특성 복사 마법사를 사용하여 데이터 파일에서 규칙을 불러올 수 있습니다.

검증 규칙 정의

검증 규칙 정의 대화 상자에서 단일-변수 및 교차-변수 검증 규칙을 작성하고 볼 수 있습니다.

검증 규칙을 작성하고 보려면 다음을 수행하십시오.

1. 메뉴에서 다음을 선택합니다.

   데이터 > 검증 > 규칙 정의...

   이 대화 상자는 데이터 사전에서 읽은 단일-변수 및 교차-변수 검증 규칙으로 채워집니다. 규칙이 없는 경우 사용자 목적에 맞게 수정할 수 있는 플레이스홀더 규칙이 자동으로 작성됩니다.
2. 단일-변수 규칙 및 교차-변수 규칙 탭에서 개별 규칙을 선택하여 해당 특성을 보고 수정할 수 있습니다.

단일-변수 규칙 정의

단일-변수 규칙 탭에서는 단일-변수 규칙을 정의하고, 보고, 수정할 수 있습니다.

규칙. 이 목록은 이름별 단일-변수 검증 규칙과 해당 규칙을 적용할 수 있는 변수 유형을 표시합니다. 이 대화 상자를 열면 데이터 사전에 정의된 규칙이 표시되거나, 현재 정의된 규칙이 없는 경우 "단일-변수 규칙 1"이라는 플레이스홀더 규칙이 표시됩니다. 다음 단추가 규칙 목록 아래에 표시됩니다.

- **새로 만들기.** 규칙 목록 맨 아래에 새 항목을 추가합니다. 규칙이 선택되고 "SingleVarRule *n*"이라는 이름으로 지정됩니다. 여기서 새 규칙의 이름이 단일-변수와 교차-변수 규칙에서 고유하도록 *n*을 정수로 지정해야 합니다.
- **중복.** 규칙 목록 맨 아래에 선택한 규칙의 사본을 추가합니다. 규칙 이름은 단일-변수와 교차-변수 규칙에서 고유하도록 조정됩니다. 예를 들어 "SingleVarRule 1"이라는 이름이 중복된 경우 처음 중복된 규칙의 이름이 "SingleVarRule 1의 사본"이 되고 두 번째가 "SingleVarRule 1의 사본 (2)" 등으로 지정됩니다.
- **삭제.** 선택한 규칙을 삭제합니다.

규칙 정의. 이러한 제어를 사용하여 선택한 규칙에 대한 특성을 보고 설정할 수 있습니다.

- **이름.** 규칙 이름은 단일-변수와 교차-변수 규칙에서 고유해야 합니다.
- **유형.** 이는 규칙을 적용할 수 있는 변수 유형입니다. **숫자**, **문자열**, **날짜**에서 선택하십시오.
- **형식.** 이를 사용하여 날짜 변수에 적용할 수 있는 규칙의 날짜 형식을 선택할 수 있습니다.
- **유효한 값.** 유효한 값을 범위 또는 값 목록으로 지정할 수 있습니다.

  범위 정의

  범위 정의 제어를 사용하여 유효한 범위를 지정할 수 있습니다. 범위 밖의 값은 유효하지 않음으로 플래그 지정됩니다.

  범위를 지정하려면 최소값 또는 최대값, 또는 둘 다를 입력하십시오. 확인 상자 컨트롤을 사용하여 범위 내에서 레이블이 지정되지 않은 비정수 값을 플래그 지정할 수 있습니다.

  목록 정의

  목록 정의 제어를 사용하여 유효한 값의 목록을 정의할 수 있습니다. 목록에 포함되지 않은 값은 유효하지 않음으로 플래그 지정됩니다.

  목록 값을 격자에 입력하십시오. 확인 상자는 케이스에서 허용 가능한 값 목록에 대해 문자열 데이터 값이 선택된 경우를 구분하는지 여부를 판별합니다.
- **사용자-결측값 허용.** 사용자-결측값을 유효하지 않음으로 플래그 지정할지 여부를 제어합니다.
- **시스템-결측값 허용.** 시스템-결측값을 유효하지 않음으로 플래그 지정할지 여부를 제어합니다. 이는 문자열 규칙 유형에는 적용되지 않습니다.
- **빈 값 허용.** 공백(완전히 비어 있음) 문자열 값을 유효하지 않음으로 플래그 지정할지 여부를 제어합니다. 이는 비문자열 규칙 유형에는 적용되지 않습니다.

교차-변수 규칙 정의

교차-변수 규칙 탭에서는 교차-변수 규칙을 정의하고, 보고, 수정할 수 있습니다.

규칙. 이 목록은 교차-변수 검증 규칙을 이름순으로 표시합니다. 이 대화 상자를 열면 "CrossVarRule 1"이라는 플레이스홀더 규칙이 표시됩니다. 다음 단추가 규칙 목록 아래에 표시됩니다.

- **새로 만들기.** 규칙 목록 맨 아래에 새 항목을 추가합니다. 규칙이 선택되고 "CrossVarRule *n*"이라는 이름으로 지정됩니다. 여기서 새 규칙의 이름이 단일-변수와 교차-변수 규칙에서 고유하도록 *n*을 정수로 지정해야 합니다.
- **중복.** 규칙 목록 맨 아래에 선택한 규칙의 사본을 추가합니다. 규칙 이름은 단일-변수와 교차-변수 규칙에서 고유하도록 조정됩니다. 예를 들어 "CrossVarRule 1"이라는 이름이 중복된 경우 처음 중복된 규칙의 이름이 "CrossVarRule 1의 사본"이 되고 두 번째가 "CrossVarRule 1의 사본 (2)" 등으로 지정됩니다.
- **삭제.** 선택한 규칙을 삭제합니다.

규칙 정의. 이러한 제어를 사용하여 선택한 규칙에 대한 특성을 보고 설정할 수 있습니다.

- **이름.** 규칙 이름은 단일-변수와 교차-변수 규칙에서 고유해야 합니다.
- **논리식.** 이는 본질적으로 규칙 정의입니다. 유효하지 않은 케이스를 1로 평가하도록 표현식을 코딩해야 합니다.

표현식 작성

1. 표현식을 작성하려면 표현식 필드에 성분을 붙여넣거나 직접 입력합니다.

- 함수 그룹 목록에서 그룹을 선택한 다음 함수 및 특수변수 목록에서 함수 또는 변수를 두 번 클릭하거나 함수 또는 변수를 선택한 다음 **삽입**을 클릭하여 함수 또는 일반적으로 사용되는 시스템 변수를 붙여넣을 수 있습니다. 물음표로 표시된 모수에 값을 입력하십시오(함수에만 적용). **모두**라는 레이블이 있는 함수 그룹에서는 사용 가능한 모든 함수와 시스템 변수의 목록을 제공합니다. 대화 상자의 지정된 영역에는 현재 선택한 함수나 변수에 대한 간략한 설명이 표시됩니다.
- 문자 상수는 따옴표나 어포스트로피로 묶어야 합니다.
- 값에 소수점이하자리수가 포함된 경우 소수점 구분자로는 마침표(.)를 사용해야 합니다.

제 3 장 데이터 검증

데이터 검증 대화 상자를 사용하여 활성 데이터 세트에서 의심이 되는 유효하지 않은 케이스, 변수 및 데이터 값을 식별할 수 있습니다.

예제. 데이터 분석가는 월별 고객 만족 보고서를 클라이언트에게 제공해야 합니다. 분석가가 매월 받는 데이터는 불완전한 고객 ID, 범위를 벗어난 변수 값, 공통적으로 오류 상태로 입력된 변수 값의 조합에 대한 품질 확인이 수행되어야 합니다. 분석가는 데이터 검증 대화 상자를 사용하여 고객을 고유하게 식별하는 변수를 지정하고, 유효한 변수 범위에 대한 단일-변수 규칙을 정의하고, 불가능한 조합을 찾는 교차-변수 규칙을 정의할 수 있습니다. 이 프로시저는 문제 케이스 및 변수 보고서를 리턴합니다. 또한 데이터는 매월 동일한 데이터 요소를 가지고 있으므로 분석가는 다음 달에도 새 데이터 파일에 해당 규칙을 적용할 수 있습니다.

통계. 이 프로시저는 여러 확인에 실패한 변수, 케이스 및 데이터 값 목록, 단일-변수 및 교차-변수 규칙 위반 수, 분석 변수의 단순 기술통계 요약을 생성합니다.

가중값. 이 프로시저는 가중값 지정 사항을 무시하고 대신 이를 다른 분석 변수로 처리합니다.

데이터를 검증하려면 다음을 수행하십시오.

1. 메뉴에서 다음을 선택합니다.

   데이터 > 검증 > 데이터 검증...

2. 기본 변수 확인 또는 단일-변수 검증 규칙을 기준으로 검증할 하나 이상의 분석 변수를 선택하십시오.

   또는 다음을 수행할 수 있습니다.

3. **교차-변수 규칙** 탭을 클릭하고 하나 이상의 교차-변수 규칙을 적용하십시오.

선택적으로 다음을 수행할 수 있습니다.

- 중복 또는 불완전 ID를 확인할 하나 이상의 케이스 식별 변수를 선택하십시오. 또한 케이스 ID 변수는 케이스별 결과의 레이블을 지정하는 데 사용됩니다. 두 개 이상의 케이스 ID 변수가 지정되면 해당 값의 조합이 케이스 식별자로 처리됩니다.

측정 수준을 알 수 없는 필드

측정 수준 경고는 데이터 세트에서 하나 이상의 변수(필드)에 대해 측정 수준을 알 수 없을 때 표시됩니다. 측정 수준은 이 프로시저의 계산 결과에 영향을 미치기 때문에 모든 변수에 정의된 측정 수준이 있어야 합니다.

데이터 스캔. 활성 데이터 세트의 데이터를 읽고 현재 알 수 없는 측정 수준이 있는 필드에 기본 측정 수준을 할당합니다. 데이터 세트가 큰 경우 시간이 걸릴 수 있습니다.

수동으로 할당. 알 수 없는 측정 수준이 있는 필드를 모두 나열하는 대화 상자를 엽니다. 이 대화 상자에서 해당 필드에 측정 수준을 할당할 수 있습니다. 데이터 편집기의 변수 보기에서도 측정 수준을 할당할 수 있습니다.

이 프로시저에 대해 측정 수준이 중요하기 때문에 모든 필드에 정의된 측정 수준이 있을 때까지는 대화 상자에 액세스하여 이 프로시저를 실행할 수 없습니다.

데이터 기본 확인 검증

기본 확인 탭을 사용하여 변수, 케이스 식별자 및 전체 식별자를 분석하기 위한 기본 확인을 선택할 수 있습니다.

분석 변수. 변수 탭에서 분석 변수를 선택한 경우 다음 유효성 확인을 선택할 수 있습니다. 확인 상자를 사용하여 확인을 설정하거나 해제할 수 있습니다.

- **결측값의 최대 퍼센트.** 결측값의 퍼센트가 지정된 값보다 큰 보고서 분석. 지정된 값은 100보다 작거나 같은 양수여야 합니다.
- **단일 범주의 최대 케이스 퍼센트.** 분석 변수가 범주형인 경우, 이 옵션은 지정된 값보다 큰 단일 비결측 범주를 나타내는 케이스 퍼센트가 있는 범주형 분석 변수를 보고합니다. 지정된 값은 100보다 작거나 같은 양수여야 합니다. 퍼센트는 변수의 결측되지 않은 값이 있는 케이스를 기반으로 합니다.
- **수가 1인 범주의 최대 퍼센트.** 분석 변수가 범주형인 경우, 이 옵션은 케이스가 하나만 있는 변수의 범주 퍼센트가 지정된 값보다 큰 범주형 분석 변수를 보고합니다. 지정된 값은 100보다 작거나 같은 양수여야 합니다.
- **최소 변동계수.** 분석 변수가 척도인 경우 이 옵션은 변동계수의 절대값이 지정된 값보다 작은 척도분석 변수를 보고합니다. 이 옵션은 평균이 0이 아닌 변수에만 적용됩니다. 지정된 값은 음수가 아니어야 합니다. 0을 지정하면 변동계수 확인이 해제됩니다.
- **최소 표준편차.** 분석 변수가 척도인 경우 이 옵션은 표준편차가 지정된 값보다 작은 척도분석 변수를 보고합니다. 지정된 값은 음수가 아니어야 합니다. 0을 지정하면 표준편차 확인을 해제합니다.

케이스 식별자. 변수 탭에서 케이스 식별자 변수를 선택한 경우 다음 유효성 확인을 선택할 수 있습니다.

- **불완전한 ID에 플래그.** 이 옵션은 불완전한 케이스 식별자가 있는 케이스를 보고합니다. 특정 케이스에 대해 ID 변수 값이 공백이거나 누락된 경우 식별자가 불완전하다고 간주됩니다.
- **중복 ID에 플래그.** 이 옵션은 중복 케이스 식별자가 있는 케이스를 보고합니다. 불완전한 식별자는 가능한 중복 세트에서 제외됩니다.

빈 케이스에 플래그. 이 옵션은 모든 변수가 비어 있거나 공백 상태인 케이스를 보고합니다. 비어 있는 케이스를 식별하기 위해 파일에서 모든 변수를 사용하도록 선택하거나(ID 변수 제외) 변수 탭에 정의된 분석 변수만 사용하도록 선택할 수 있습니다.

데이터 검증 단일-변수 규칙

단일-변수 규칙 탭은 사용 가능한 단일-변수 검증 규칙을 표시하고 이 탭에서 해당 단일-변수 검증 규칙을 사용하여 변수를 분석할 수 있습니다. 추가 단일-변수 규칙을 정의하려면 **규칙 정의**를 클릭하십시오. 자세한 정보는 3 페이지의 『단일-변수 규칙 정의』 주제를 참조하십시오.

분석 변수. 이 목록은 분석 변수를 표시하고, 해당 분포를 요약하며, 각 변수에 적용되는 규칙 수를 표시합니다. 이 요약에는 사용자 및 시스템 결측값이 포함되지 않음을 참고하십시오. 표시 드롭 다운 목록은 표시되는 변수를 제어합니다. **모든 변수**, **숫자변수**, **문자변수** 및 **날짜변수**에서 선택할 수 있습니다.

규칙. 분석 변수에 규칙을 적용하려면 하나 이상의 변수를 선택하고 규칙 목록에서 적용할 모든 규칙을 선택하십시오. 규칙 목록은 선택된 분석 변수에 적합한 규칙만 표시합니다. 예를 들어 숫자 분석 변수를 선택한 경우 숫자 규칙만 표시되고 문자변수를 선택한 경우 문자열 규칙만 표시됩니다. 분석 변수가 선택되지 않았거나 혼합된 데이터 유형이 있는 경우 규칙이 표시되지 않습니다.

변수 분포. 분석 변수 목록에 표시된 분포 요약은 케이스 입력란에 지정된 대로 모든 케이스를 기준으로 하거나 처음 *n*개의 케이스로 제한됩니다. **재스캔**을 클릭하면 분포 요약 정보가 업데이트됩니다.

검증 데이터 교차-변수 규칙

교차-변수 규칙 탭은 사용 가능한 교차-변수 규칙을 표시하고, 여기서 해당 규칙을 사용자 데이터에 적용할 수 있습니다. 추가 교차-변수 규칙을 정의하려면 **규칙 정의**를 클릭하십시오. 자세한 정보는 4 페이지의 『교차-변수 규칙 정의』 주제를 참조하십시오.

데이터 검증 결과

케이스별 보고서. 단일-변수 또는 교차-변수 검증 규칙을 적용한 경우 개별 케이스에 대한 검증 규칙 위반을 나열하는 보고서를 요청할 수 있습니다.

- **최소 위반 수.** 보고서에 포함시키기 위해 케이스에 필요한 최소 규칙 위반 수를 지정합니다. 양의 정수를 지정합니다.
- **최대 케이스 수.** 이 옵션은 케이스 보고서에 포함된 최대 케이스 수를 지정합니다. 1000보다 작거나 같은 양의 정수를 지정하십시오.

단일-변수 검증 규칙. 단일-변수 검증 규칙을 적용한 경우 결과 표시 방법과 결과를 모두 표시할지 여부를 선택할 수 있습니다.

- **분석 변수별로 위반 요약.** 각 분석 변수에 대해 이 옵션은 위반된 모든 단일-변수 검증 규칙과 각 규칙을 위반한 값 수를 표시합니다. 또한 각 변수에 대한 총 단일-변수 규칙 위반 수를 보고합니다.
- **규칙별로 위반 요약.** 각 단일-변수 검증 규칙에 대해 이 옵션은 규칙을 위반한 변수를 보고하고 변수당 유효하지 않은 값 수를 보고합니다. 또한 변수에서 각 규칙을 위반한 총 값 수를 보고합니다.

분석 변수에 대한 기술통계량 표시. 이 옵션을 사용하여 분석 변수에 대한 기술통계량을 요청할 수 있습니다. 각 범주형 변수에 대해 빈도표가 생성됩니다. 척도변수에 대해 평균, 표준편차, 최소값 및 최대값이 포함된 요약 통계량 표가 생성됩니다.

검증 규칙 위반이 있는 케이스를 활성 데이터 세트의 맨 위로 이동. 이 옵션은 단일-변수 또는 교차-변수 규칙 위반이 있는 케이스를 쉽게 볼 수 있도록 활성 데이터 세트의 맨 위로 이동시킵니다.

데이터 검증 저장

저장 탭을 사용하여 활성 데이터 세트에 규칙 위반을 기록하는 변수를 저장할 수 있습니다.

요약변수. 저장할 수 있는 개별 변수가 있습니다. 변수를 저장하려면 상자를 선택하십시오. 변수의 기본 이름이 제공됩니다. 이를 편집할 수 있습니다.

- **빈 케이스 표시자.** 빈 케이스는 값 1로 지정됩니다. 다른 모든 케이스는 0으로 코딩됩니다. 변수 값에는 기본 확인 탭에서 지정된 점수가 반영됩니다.
- **중복 ID 집단.** 케이스 식별자가 동일한 케이스(불완전한 식별자가 있는 케이스가 아님)는 동일한 집단 번호로 지정됩니다. 고유하거나 불완전한 식별자의 케이스는 0으로 코딩됩니다.
- **불완전한 ID 표시자.** 비어 있거나 불완전한 케이스 식별자가 있는 케이스는 값 1로 지정됩니다. 다른 모든 케이스는 0으로 코딩됩니다.
- **검증 규칙 위반.** 이는 케이스별 단일-변수 및 교차-변수 검증 규칙 위반의 총 수입니다.

기존 요약변수 바꾸기. 데이터 파일에 저장된 변수는 고유한 이름을 가지고 있거나 동일한 이름의 변수로 대체해야 합니다.

표시자 변수 저장. 이 옵션을 사용하여 검증 규칙 위반의 완전한 레코드를 저장할 수 있습니다. 각 변수는 검증 규칙의 애플리케이션에 해당하고, 케이스가 규칙을 위반하는 경우 값이 1이며 위반하지 않는 경우 값이 0입니다.

제 4 장 자동 데이터 준비

분석할 데이터 준비는 모든 프로젝트에서 가장 중요한 단계 중 하나이며 일반적으로 가장 많은 시간이 걸리는 단계입니다. 자동 데이터 준비(ADP)는 데이터 분석 및 수정사항 식별, 문제점이 있거나 유용하지 않은 필드 선별, 해당하는 경우 새 속성 파생, 지능적 선별 기법을 통한 성능 개선 작업과 같이 사용자를 위한 작업을 처리합니다. 수정사항을 선택하여 적용할 수 있는 완전히 **자동화**된 방법으로 알고리즘을 사용하거나, 변경하기 전에 변경사항을 미리 보고 원하는 대로 수용하거나 거부할 수 있는 **대화형** 방법으로 알고리즘을 사용할 수 있습니다.

통계 개념의 고급 지식을 습득하지 않고도 ADP를 사용하여 모형설정을 위한 데이터 준비를 신속하고 쉽게 처리할 수 있습니다. 모형은 더 신속하게 작성되고 점수화되며, 또한 ADP를 사용하여 자동 모형화 프로세스의 견고함을 향상시킵니다.

참고: ADP에서 분석을 위해 필드를 준비하는 경우 이전 필드의 기존 값과 특성을 바꾸지 않고 조정되거나 변환된 새 필드를 작성합니다. 상세 분석에서는 이전 필드가 사용되지 않으며 해당 역할이 없음으로 설정되어 있습니다. 또한 모든 사용자 결측값 정보가 새로 작성된 필드로 전송되지 않으며 새 필드의 결측값은 시스템 결측값입니다.

예제. 집 소유자의 보험 청구를 조사하기 위한 자원이 한정된 보험 회사는 의심되는 잠재적인 사기 청구에 대한 모형을 작성하려고 합니다. 모형을 작성하기 전에 자동 데이터 준비를 사용하여 모형화를 위한 데이터를 준비합니다. 변환을 적용하기 전에 제안된 변환을 검토할 수 있도록 하기 위해 자동 데이터 준비를 대화식 모드로 사용합니다.

자동화된 산업 그룹에서는 여러 종류의 개인용 차량 판매를 계속해서 추적합니다. 실적이 좋은 모형과 실적이 저조한 모형을 식별하기 위해 차량 판매와 차량 특성 간의 관계를 구축하려고 합니다. 자동 데이터 준비를 사용하여 분석을 위한 데이터를 준비하고, 결과가 어떻게 다른지를 보기 위해 준비 "전"과 "후"에 데이터를 사용하여 모형을 작성합니다.

원하는 목적. 자동 데이터 준비에서는 모형을 작성하고 해당 모형의 예측력을 향상시킬 수 있는 다른 알고리즘을 사용하여 속도에 영향을 주는 데이터 준비 단계를 권장합니다. 여기에는 변환, 구성 및 선택 기능이 포함될 수 있습니다. 또한 목표도 변환될 수 있습니다. 데이터 준비 프로세스에서 집중해야 하는 모형설정 우선 순위를 지정할 수 있습니다.

- **속도 및 정확도 균형.** 이 옵션은 모형설정 알고리즘이 데이터를 처리하는 속도와 예측의 정확도 둘 다에 동일한 우선 순위를 부여하도록 데이터를 준비합니다.
- **속도 최적화.** 이 옵션은 모형설정 알고리즘이 데이터를 처리하는 속도에 우선 순위를 부여하도록 데이터를 준비합니다. 매우 큰 데이터 세트를 사용하여 작업하거나 빠른 해답을 찾으려는 경우 이 옵션을 선택하십시오.
- **정확도 최적화.** 이 옵션은 모형설정 알고리즘으로 생성되는 예측의 정확도에 우선 순위를 부여하도록 데이터를 준비합니다.

- **사용자 정의 분석.** 설정 탭에서 알고리즘을 수동으로 수정할 경우 이 옵션을 선택하십시오. 이후에 설정 탭에서 다른 목적 중 하나와 호환되지 않는 옵션을 변경하는 경우 이 설정이 자동으로 선택됨을 참고하십시오.

자동 데이터 준비 확보

메뉴에서 다음을 선택합니다.

1. 메뉴에서 다음을 선택합니다.

   변환 > 모형화를 위한 데이터 준비 > 자동...

2. **실행**을 클릭합니다.

선택적으로 다음을 수행할 수 있습니다.

- 목적 탭에서 목적을 지정합니다.
- 필드 탭에서 필드 할당을 지정합니다.
- 설정 탭에서 전문가 설정을 지정합니다.

대화형 데이터 준비 확보

1. 메뉴에서 다음을 선택합니다.

   변환 > 모형화를 위한 데이터 준비 > 대화형...

2. 대화 상단의 맨 위에 있는 도구 모음에서 **분석**을 클릭하십시오.
3. 분석 탭을 클릭하여 제안된 데이터 준비 단계를 검토하십시오.
4. 충족하는 경우 **실행**을 클릭하십시오. 그렇지 않으면 **분석 지우기**를 클릭하여 원하는 설정을 변경하고 **분석**을 클릭하십시오.

선택적으로 다음을 수행할 수 있습니다.

- 목적 탭에서 목적을 지정합니다.
- 필드 탭에서 필드 할당을 지정합니다.
- 설정 탭에서 전문가 설정을 지정합니다.
- **XML 저장**을 클릭하여 제안된 데이터 준비 단계를 XML 파일로 저장합니다.

필드 탭

이 필드 탭에서는 추가 분석을 위해 준비해야 할 필드를 지정합니다.

사전 정의된 역할 사용. 이 옵션을 기존 필드 정보를 사용합니다. 역할이 목표인 단일 필드가 있는 경우 이는 목표로 사용되고, 그렇지 않은 경우 목표가 없습니다. 역할이 입력으로 사전 정의된 모든 필드는 입력으로 사용됩니다. 하나 이상의 입력 필드가 있어야 합니다.

사용자 정의 필드 할당 사용. 기본 목록에서 필드를 이동하여 필드 역할을 대체하는 경우 대화 상자에서는 자동으로 해당 옵션으로 전환됩니다. 사용자 정의 필드 지정을 작성할 때 다음 필드를 지정하십시오.

- **대상(선택적).** 목표가 필요한 모형을 작성할 경우 목표 필드를 선택하십시오. 이는 필드 역할을 목표로 설정하는 것과 유사합니다.
- **입력.** 하나 이상의 입력 필드를 선택합니다. 이는 필드 역할을 입력으로 설정하는 것과 유사합니다.

설정 탭

설정 탭은 몇 가지 다른 설정 그룹으로 구성되어 있으며 설정을 수정하여 알고리즘이 데이터를 처리하는 방법을 미세 튜닝할 수 있습니다. 기본값 설정을 다른 목적과 함께 사용할 수 없는 사항으로 변경하면, 목적 탭이 자동으로 업데이트되어 **분석 사용자 정의** 옵션이 선택됩니다.

날짜 및 시간 준비

많은 모형화 알고리즘에서는 날짜 및 시간 세부사항을 직접 처리할 수 없습니다. 이러한 설정을 사용하면 기존 데이터의 날짜 및 시간에서 모형 입력으로 사용할 수 있는 새 기간 데이터를 파생시킬 수 있습니다. 날짜 및 시간이 포함된 필드는 날짜 또는 시간 저장 공간 유형을 사용하여 사전 정의되어야 합니다. 원래 날짜 및 시간 필드는 자동 데이터 준비를 따르는 모형 입력으로 권장되지 않습니다.

모형화 날짜 및 시간 준비. 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 날짜 및 시간 준비 제어를 사용 안함으로 설정합니다.

참조 날짜까지의 경과 시간 계산. 이는 날짜가 포함된 각 변수의 참조 날짜 이후의 년/월/일 수를 생성합니다.

- **참조 날짜.** 입력 데이터의 날짜 정보와 관련하여 계산된 기간으로부터 날짜를 지정합니다. **오늘 날짜**를 선택하면 ADP가 실행될 때 항상 현재 시스템 날짜가 사용됩니다. 특정 날짜를 사용하려면 **고정 날짜**를 선택하고 필요한 날짜를 입력하십시오.
- **날짜 기간 단위.** ADP가 날짜 기간 단위를 자동으로 결정해야 하는지 여부를 지정하거나 **고정 단위**를 년, 월 또는 일로 선택하십시오.

참조 시간까지의 경과 시간 계산. 이는 날짜가 포함된 각 변수의 참조 시간 이후의 시/분/초를 생성합니다.

- **참조 시간.** 입력 데이터의 시간 정보와 관련하여 계산된 기간으로부터 시간을 지정합니다. **현재 시간**을 선택하면 ADP가 실행될 때 항상 현재 시스템 시간이 사용됩니다. 특정 시간을 사용하려면 **고정 시간**을 선택하고 필요한 세부사항을 입력하십시오.
- **시간 기간 단위.** ADP가 시간 기간 단위를 자동으로 결정해야 하는지 여부를 지정하거나 **고정 단위**를 시, 분 또는 초로 선택하십시오.

순환 시간 요소 추출. 이 설정을 사용하여 단일 날짜 또는 시간을 하나 이상의 필드로 분할합니다. 예를 들어 세 개의 날짜 확인 상자를 모두 선택한 경우 입력 날짜 필드 "1954-05-23"이 **필드 이름** 패널에 정의된 접미 문자를 각각 사용하여 1954, 5, 23이라는 세 개의 필드로 분할되고 원래 날짜 필드는 무시됩니다.

- **날짜에서 추출.** 날짜 입력의 경우 년, 월, 일 또는 조합을 추출할지 여부를 지정합니다.

- **시간에서 추출.** 날짜 입력의 경우 시, 분, 초 또는 조합을 추출할지 여부를 지정합니다.

필드 제외

품질이 나쁜 데이터는 예측 정확도에 영향을 미칠 수 있으므로 입력 기능에 허용 가능한 품질 수준을 지정할 수 있습니다. 상수이거나 결측값이 100%인 모든 필드는 자동으로 제외됩니다.

저품질 입력 필드 제외. 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 제외 필드 제어를 사용 안함으로 설정합니다.

결측값이 너무 많은 필드 제외. 지정된 결측값 퍼센트보다 큰 필드가 추가 분석에서 제거됩니다. 0(이 옵션을 선택 취소함)보다 크거나 같고 100(모든 결측값이 있는 필드를 자동으로 제외)보다 작거나 같은 값을 지정하십시오. 기본값은 50입니다.

고유 범주가 너무 많은 명목형 필드 제외. 지정된 수보다 범주의 수가 많은 필드가 추가 분석에서 제거됩니다. 양의 정수를 지정합니다. 기본값은 100입니다. 이는 모형화에서 레코드 고유 정보(예: ID, 주소 또는 이름)가 포함된 필드를 자동으로 제거하는 경우에 유용합니다.

단일 범주에 값이 너무 많은 범주형 필드 제외. 지정된 퍼센트보다 큰 레코드가 포함된 범주가 있는 순서형 및 명목형 필드가 추가 분석에서 제거됩니다. 0(이 옵션을 선택 취소함)보다 크거나 같고 100(상수 필드가 자동으로 제외됨)보다 작거나 같은 값을 지정하십시오. 기본값은 95입니다.

측정 수정

측정 수준 수정. 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 측정 수정 제어를 사용 안함으로 설정합니다.

측정 수준. 값이 "너무 적음"인 연속형 필드의 측정 수준을 순서형으로 조정할 수 있는지와 값이 "너무 많음"인 순서형 필드를 연속형으로 수정할 수 있는지 여부를 지정합니다.

- **순서형 필드 값의 최대 수.** 지정된 범주 수보다 많은 순서형 필드는 연속형 필드로 형변환됩니다. 양의 정수를 지정합니다. 기본값은 10입니다. 이 값은 연속형 필드 값의 최소 수보다 크거나 같아야 합니다.
- **연속형 필드 값의 최소 수.** 지정된 고유 값 수보다 적은 연속형 필드는 순서형 필드로 형변환됩니다. 양의 정수를 지정합니다. 기본값은 5입니다. 이 값은 순서형 필드 값의 최대 수보다 작거나 같아야 합니다.

데이터 품질 향상

데이터 품질 향상을 위해 필드 준비. 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 데이터 품질 향상 제어를 사용 안함으로 설정합니다.

이상값 처리. 입력 및 목표에 대한 이상값을 대체할지 여부를 지정합니다. 대체하는 경우 이상값 분리점 기준, 표준편차의 측정값 및 이상값 대체 방법을 지정하십시오. 제거(분리점 값으로 설정)하거나 이상값을 결측값으로 설정하여 이상값을 대체할 수 있습니다. 결측으로 설정 이상값은 아래에 선택된 결측값 처리 설정을 따릅니다.

결측값 대체. 연속형, 명목형 또는 순서형 필드의 결측값을 대체할지 여부를 지정합니다.

명목형 필드 순서 바꾸기. 명목형(세트) 필드 값을 가장 작은 값(최소 빈도 발생)에서 가장 큰 값(최대 빈도 발생)의 범주로 기록하려면 이를 선택하십시오. 새 필드 값은 가장 작은 빈도 범주로 0부터 시작합니다. 원래 필드가 문자열인 경우에도 새 필드는 숫자입니다. 예를 들어 명목형 필드의 데이터 값이 "A", "A", "A", "B", "C", "C"인 경우, 자동 데이터 준비에서는 "B"를 0으로, "C"를 1로, "A"를 2로 기록합니다.

필드 조정

필드 조정. 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 필드 조정 제어를 사용 안함으로 설정합니다.

분석 가중치. 이 변수에는 분석(회귀분석 또는 표본추출) 가중치가 포함됩니다. 계정에 분석 가중값을 사용하여 목표 필드의 수준 간의 분산 차이를 고려합니다. 연속형 필드를 선택하십시오.

연속형 입력 필드. 이는 **z-점수 변환** 또는 **최대/최소 변환**을 사용하여 연속형 입력을 표준화합니다. 선택 및 구성 설정에서 **변수 구성 수행**을 선택한 경우 특히 입력 조정이 유용합니다.

- **Z-점수 변환.** 관측 평균 및 표준편차를 모수 추정값 채우기로 사용하면 필드가 표준화되고 *z* 점수가 **최종 평균** 및 **최종 표준 편차**가 지정된 정규 분포의 해당 값으로 매핑됩니다. **최종 평균**에 수를 지정하고 **최종 표준편차**에 양수를 지정하십시오. 기본값은 각각 표준화된 조정값인 0과 1입니다.
- **최대/최소 변환.** 관측된 최대값 및 최소값을 모수 추정값 채우기로 사용하면 **최소값** 및 **최대값**이 지정된 균일 분포의 해당 값으로 필드가 매핑됩니다. **최대값**이 **최소값**보다 큰 수를 지정하십시오.

연속형 목표. 이는 Box-Cox 변환을 사용하는 연속형 목표를 **최종 평균** 및 **최종 표준편차**가 지정된 대략적인 정규 분포가 있는 필드로 변환합니다. **최종 평균**에 수를 지정하고 **최종 표준편차**에 양수를 지정하십시오. 기본값은 각각 0과 1입니다.

참고: ADP에서 목표를 변환한 경우 변환된 목표를 사용하여 작성된 이후의 모형은 변환된 단위로 점수가 지정됩니다. 결과를 해석하고 사용하려면 예측값을 다시 원래 척도로 변환해야 합니다. 자세한 정보는 주제를 참조하십시오. 자세한 정보는 25 페이지의 『역변환 점수』 주제를 참조하십시오.

필드 변환

데이터의 예측력을 향상시키기 위해 입력 필드를 변환할 수 있습니다.

모형화를 위한 필드 변환. 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 필드 변환 제어를 사용 안함으로 설정합니다.

범주형 입력 필드 다음 옵션을 사용할 수 있습니다.

- **목표와의 연관성을 극대화하도록 희박한 범주 병합.** 이를 선택하면 목표와 연관하여 처리되는 필드 수를 줄여 더욱 경제적인 모형을 만들 수 있습니다. 입력과 대상 간의 관계를 기반으로 유사한 범주가 식별됩니다. 유의적으로 다르지 않은 범주(즉 지정된 값보다 큰 *p*-값을 가지는 범주)가 합쳐집니다. 0보다 크고 1보다 작거나 같은 값을 지정하십시오. 모든 범주가 하나로 합쳐지는 경우, 필드의 원래 버전과 파생된 버전이 예측변수로서 값이 없기 때문에 추가 분석에서 제외됩니다.

- **목표가 없는 경우 수를 기준으로 희박한 범주 병합.** 데이터 세트에 목표가 없는 경우 순서형 및 명목형 필드의 희박한 범주를 병합하도록 선택할 수 있습니다. 총 레코드 수에 대해 지정된 최소 퍼센트보다 작은 범주와 병합하는 경우 동일 빈도 방법이 사용됩니다. 0보다 크거나 같고 100보다 작거나 같은 값을 지정합니다. 기본값은 10입니다. 케이스에 지정된 최소 퍼센트보다 작은 범주가 없거나 두 개의 범주만 남아 있는 경우 병합이 중지됩니다.

연속형 입력 필드. 데이터 세트에 범주형 목표가 없는 경우 연관성이 더 강한 연속형 입력을 구간화하여 처리 성능을 개선할 수 있습니다. 구간화는 "동질적 부분집합" 특성을 기반으로 작성되는데, 이는 지정된 *p* 값을 동질적 부분집합을 판별하기 위한 임계값의 알파로 사용하는 Scheffe 방법으로 식별됩니다. 0보다 크고 1보다 작거나 같은 값을 지정하십시오. 기본값은 0.05입니다. 구간화 작업을 통해 특정 필드에 대한 단일 구간이 생성된 경우 예측자로 사용할 값이 없으므로 원본 및 구간화된 필드 버전이 제외됩니다.

참고: ADP의 구간화는 최적 구간화와 다릅니다. 최적 구간화는 엔트로피 정보를 사용하여 연속형 필드를 범주형 필드로 변환합니다. 이를 수행하려면 데이터를 정렬하고 이를 모두 메모리에 저장해야 합니다. ADP는 동질적 부분집합을 사용하여 연속형 필드를 구간화하는데, 이는 ADP 구간화에서 날짜를 정렬하고 모든 데이터를 메모리에 저장할 필요가 없음을 의미합니다. 동질적 부분집합 방법을 사용하여 연속형 필드를 구간화하는 경우 구간화 후의 범주 수는 항상 목표의 범주 수보다 작거나 같습니다.

선택 및 구성

데이터의 예측력을 향상시키기 위해 기존 필드를 기반으로 하여 새 필드를 구성할 수 있습니다.

변수 선택 수행. 목표의 상관관계에 대한 *p* 값이 지정된 *p* 값보다 큰 경우 분석에서 연속 입력이 제거됩니다.

변수 구성 수행. 이 옵션을 선택하여 몇 가지 기존 변수의 조합에서 새 변수를 파생시킵니다. 이전 변수는 추가 분석에서 사용되지 않습니다. 이 옵션은 목표가 연속적이거나 목표가 없는 연속 입력 변수에만 적용됩니다.

필드 이름

새 기능 및 변환된 기능을 쉽게 식별하기 위해 ADP에서 기본 새 이름, 접두문자 또는 접미문자를 작성하고 적용합니다. 사용자 요구 및 데이터와 더 관련되도록 해당 이름을 수정할 수 있습니다.

변환 및 구성된 필드. 변환된 대상 및 입력 필드에 적용할 이름 확장자를 지정하십시오.

또한 선택 및 구성 설정을 통해 구성된 기능에 적용할 접두문자 이름을 지정하십시오. 이 접두문자 루트 이름에 숫자 접미문자를 추가하여 새 이름이 작성됩니다. 숫자 형식은 새 기능이 파생된 수에 따라 다릅니다. 예를 들어 다음과 같습니다.

- 1-9개의 구성된 기능 이름은 feature1 - feature9입니다.
- 10-99개의 구성된 기능 이름은 feature01 - feature99입니다.
- 100-999개의 구성된 기능 이름은 feature001 - feature999 등입니다.

이렇게 하면 구성된 기능이 파생된 수에 상관 없이 민감한 순서로 정렬됩니다.

날짜 및 시간에서 계산된 기간. 날짜 및 시간에서 계산된 기간에 적용할 이름 확장자를 지정하십시오.

날짜 및 시간에서 추출된 순환 요소. 날짜 및 시간에서 추출된 순환 요소에 적용할 이름 확장자를 지정하십시오.

변환 적용 및 저장

대화형 또는 자동 데이터 준비를 사용하는지 여부에 따라 변환 적용 및 저장에 대한 설정이 약간 다릅니다.

대화형 데이터 준비 변환 적용 설정

변환된 데이터. 이러한 설정은 변환된 데이터를 저장할 위치를 지정합니다.

- **활성 데이터 세트에 새 필드 추가.** 자동 데이터 준비로 작성된 필드가 활성 데이터 세트에 새 필드로 추가됩니다. **분석된 필드의 역할 업데이트**는 자동 데이터 준비에 의해 추가 분석에서 제외된 필드에 대한 역할을 없음으로 설정합니다.
- **새 데이터 세트 작성 또는 변환된 데이터가 있는 파일 작성.** 자동 데이터 준비에서 권장하는 필드가 새 데이터 세트 또는 파일에 추가됩니다. **분석되지 않은 필드 포함**은 필드 탭에서 지정되지 않은 원래 데이터 세트의 필드를 새 데이터 세트에 추가합니다. 이는 모형화하는 데 사용되지 않은 정보(예: ID, 주소 또는 이름)가 포함된 필드를 새 데이터 세트로 전송하는 데 유용합니다.

자동 데이터 준비 적용 및 저장 설정

변환된 데이터 그룹은 대화형 데이터 준비에서와 동일합니다. 자동 데이터 준비에는 다음 추가 옵션이 있습니다.

변환 적용. 자동 데이터 준비 대화 상자에서 이 옵션을 선택 취소하면 선택사항을 유지보수하는 동안 다른 모든 적용 및 저장 제어를 사용 안함으로 설정합니다.

변환을 명령문으로 저장. 이는 권장된 변환을 명령 구문으로 외부 파일에 저장합니다. 대화형 데이터 준비 대화 상자에서는 **붙여넣기**를 클릭하여 변환을 명령 구문으로 구문 창에 붙여넣기 때문에 이 제어가 없습니다.

변환을 XML로 저장. 이는 권장되는 변환을 XML로 외부 파일에 저장합니다. 이 외부 파일은 TMS MERGE를 사용하여 모형 PMNL에 병합되거나 TMS IMPORT를 사용하여 다른 데이터 세트에 적용될 수 있습니다. 대화형 데이터 준비 대화 상자에서는 대화 상자의 맨 위에 있는 **XML 저장**을 클릭하여 변환을 XML로 저장하기 때문에 이 제어가 없습니다.

분석 탭

참고: 대화형 데이터 준비 대화 상자에서 분석 탭을 사용하여 권장되는 변환을 검토할 수 있습니다. 자동 데이터 준비 대화 상자에는 이 단계가 없습니다.

1. 목적, 필드 및 설정 탭에서의 변경사항을 포함하여 ADP 설정에 만족하는 경우 **데이터 분석**을 클릭하십시오. 알고리즘에서 데이터 입력에 대한 설정을 적용하고 분석 탭에 결과를 표시합니다.

분석 탭에는 데이터 처리 과정이 요약되고 점수화를 위해 데이터를 수정하거나 개선할 수 있는 방법에 대한 권장사항을 표시하는 표 형식 그래프 형식의 결과 모두가 포함됩니다. 그런 다음 해당 권장사항을 검토하고 승인하거나 거부할 수 있습니다.

분석 탭은 2개 패널 즉, 왼쪽에 주 보기와 오른쪽에 링크 또는 보조 보기로 구성되어 있습니다. 세 개의 주 보기가 있습니다.

- 필드 처리 요약(기본값). 자세한 정보는 『필드 처리 요약』 주제를 참조하십시오.
- 필드. 자세한 정보는 19 페이지의 『필드』 주제를 참조하십시오.
- 동작 요약. 자세한 정보는 20 페이지의 『동작 요약』 주제를 참조하십시오.

4개의 링크/보조 보기가 있습니다.

- 예측력(기본값). 자세한 정보는 20 페이지의 『예측력』 주제를 참조하십시오.
- 필드 표. 자세한 정보는 21 페이지의 『필드 표』 주제를 참조하십시오.
- 필드 세부사항. 자세한 정보는 21 페이지의 『필드 세부사항』 주제를 참조하십시오.
- 동작 세부사항. 자세한 정보는 22 페이지의 『동작 세부사항』 주제를 참조하십시오.

보기 간 링크

주 보기 내의 표에서 밑줄이 있는 텍스트는 링크 보기 표시를 제어합니다. 텍스트를 클릭하면 특정 필드, 필드 세트 또는 처리 단계에 대한 세부사항이 표시됩니다. 마지막으로 선택한 링크에 더 진한 색상이 표시됩니다. 이를 통해 두 개의 보기 패널의 내용 사이의 연결을 식별할 수 있습니다.

보기 재설정

원래 분석 권장사항을 다시 표시하고 분석 보기에 수정한 사항을 취소하려면 주 보기 패널의 아래에 있는 **재설정**을 클릭하십시오.

필드 처리 요약

필드 처리 요약 표에서는 기능의 상태와 구성된 기능 수에 대한 변경사항을 포함하여 투영된 전체 처리 영향의 스냅샷을 제공합니다.

실제로 모형이 작성되지 않으므로 데이터 준비 전 후에 전체 예측력의 변경에 대한 측정 또는 그래프가 없습니다. 대신 개별 권장 예측자의 예측력에 대한 그래프가 표시됩니다.

이 표는 다음 정보를 표시합니다.

- 목표 필드의 수.
- 원래(입력) 예측자의 수.
- 분석 및 모형화에 사용하도록 권장된 예측자. 여기에는 다음과 같은 권장되는 총 필드 수가 포함됩니다. 권장된 원본, 변형, 필드 수, 권장되는 변형된 필드 수(필드의 중간 버전, 날짜/시간 예측자에서 파생된 필드, 구성된 예측자 제외), 날짜/시간 필드에서 파생된 권장된 필드 수, 권장된 구성된 예측자 수.

- 어떤 양식에도 권장되지 않는 입력 예측자의 수(원래 양식에서 파생된 필드 또는 구성된 예측자에 대한 입력인지 여부).

필드 정보에 밑줄이 있는 경우 이를 클릭하면 추가 세부사항이 링크된 보기로 표시됩니다. **목표, 입력 기능** 및 **사용되지 않는 입력 기능**의 세부사항이 필드 표 링크 보기에 표시됩니다. 자세한 정보는 21 페이지의 『필드 표』 주제를 참조하십시오. **분석에 사용하도록 권장된 기능**은 예측력 링크 보기에 표시됩니다. 자세한 정보는 20 페이지의 『예측력』 주제를 참조하십시오.

필드

필드 주 보기는 처리된 필드와 ADP가 다운스트림 모형에서 해당 필드를 사용하도록 권장할지 여부를 표시합니다. 필드의 권장사항은 대체할 수 있습니다. 예를 들어 구성된 기능을 제외하거나 ADP에서 제외하도록 권장한 기능을 포함시킬 수 있습니다. 필드가 변환되면 제안된 변환을 승인할지 또는 원래 버전을 사용할지 여부를 결정할 수 있습니다.

필드 보기는 두 개의 표로 구성되어 있는데, 하나는 목표에 대한 표이고 하나는 처리되었거나 작성된 예측자에 대한 표입니다.

목표 표

목표 표는 데이터에 목표가 정의된 경우에만 표시됩니다.

이 표에는 두 개의 열이 있습니다.

- **이름.** 이는 목표 필드의 이름 또는 레이블입니다. 필드가 변환되어도 원래 이름이 항상 사용됩니다.
- **측정 수준.** 이는 측정 수준을 나타내는 아이콘을 표시합니다. 아이콘 위에 마우스를 올리면 데이터를 설명하는 레이블(연속, 순서, 명목 등)이 표시됩니다.

  목표가 변환된 경우 **측정 수준** 열이 변환된 최종 버전을 반영합니다. *참고*: 목표에 대한 변환을 해제할 수 없습니다.

예측자 표

예측자 표는 항상 표시됩니다. 표의 각 행은 필드를 나타냅니다. 기본적으로 행은 예측력의 내림차순으로 정렬됩니다.

일반적인 기능의 경우 항상 원래 이름이 행 이름으로 사용됩니다. 날짜/시간 필드의 원래 및 파생 버전 둘 다 표에 표시됩니다(별도의 행으로). 또한 표에는 구성된 예측자도 포함됩니다.

표에 표시된 필드의 변환된 버전은 항상 최종 버전을 나타냄을 참고하십시오.

기본적으로 예측자 표에서는 권장 필드만 표시됩니다. 나머지 필드를 표시하려면 표의 위에 있는 **표에 비권장 필드 포함** 상자를 선택하십시오. 해당 필드가 표의 아래에 표시됩니다.

표에 다음 열이 포함됩니다.

- **사용할 버전.** 이는 필드가 다운스트림에 사용될지와 제안된 변환을 사용할지 여부를 제어하는 드롭 다운 목록을 표시합니다. 기본적으로 드롭 다운 목록은 권장사항을 반영합니다.

  변환된 일반 예측자의 경우 이 드롭 다운 목록에는 다음 세 가지 선택사항이 있습니다. **변환**, **원본** 및 **사용하지 않음**.

  변환되지 않은 경우 일반 예측자의 경우 선택사항은 **원본** 및 **사용하지 않음**입니다.

  파생된 날짜/시간 필드 및 구성된 예측자의 경우 선택사항은 **변환** 및 **사용하지 않음**입니다.

  원래 날짜/필드의 경우 드롭 다운 목록을 사용할 수 없으며 **사용하지 않음**으로 설정됩니다.

  참고: 원본 및 변환 버전이 둘 다 있는 예측자의 경우, **원본**과 **변환** 버전 간을 변경하면 해당 기능에 대한 **측정 수준** 및 **예측력** 설정이 자동으로 업데이트됩니다.
- **이름.** 각 필드 이름이 링크로 표시됩니다. 이름을 클릭하면 링크된 보기에서 필드에 대한 추가 정보를 표시합니다. 자세한 정보는 21 페이지의 『필드 세부사항』 주제를 참조하십시오.
- **측정 수준.** 이는 데이터 유형을 나타내는 아이콘을 표시합니다. 아이콘 위에 마우스를 올리면 데이터를 설명하는 레이블(연속, 순서, 명목 등)이 표시됩니다.
- **예측력.** 예측력은 ADP에서 권장하는 필드에 대해서만 표시됩니다. 목표가 없는 경우에는 이 열이 표시되지 않습니다. 예측력의 범위는 0 - 1 사이이며 값이 클 수록 '더 나은' 예측자를 나타냅니다. 일반적으로 예측력은 ADP 분석 내에서 예측자를 비교하는 데 유용하지만 분석에서 예측력 값을 비교할 수 없습니다.

동작 요약

자동 데이터 준비로 수행된 각 동작의 경우 입력 예측자가 변환 및/또는 필터링 아웃됩니다. 한 동작이 남은 필드는 다음에 사용됩니다. 마지막 단계까지 남은 필드는 모형화에 사용하도록 권장됩니다. 반면 변환 및 구성된 예측자에 대한 입력은 필터링 아웃됩니다.

동작 요약은 ADP에서 수행한 처리 동작을 나열하는 단순 표입니다. **동작**에 밑줄이 있는 경우 이를 클릭하면 수행된 동작에 대한 추가 세부사항이 링크 보기에 표시됩니다. 자세한 정보는 22 페이지의 『동작 세부사항』 주제를 참조하십시오.

참고: 각 필드의 원래 및 최종 변환된 버전만 표시되며 분석 중에 사용된 중간 버전은 표시되지 않습니다.

예측력

분석이 처음 실행되거나 필드 처리 요약 주 보기에서 **분석에 사용하도록 권장된 예측자**를 선택한 경우 이 도표에 권장된 예측자의 예측력이 표시됩니다. 필드는 예측력을 기준으로 정렬되는데 큰 값을 가진 필드가 상위에 표시됩니다.

변환된 버전의 일반 예측자의 경우 설정 탭의 필드 이름 패널에서 필드 이름은 사용자가 선택한 접두문자를 반영합니다(예: *_transformed*).

측정 수준 아이콘이 개별 필드 이름 뒤에 표시됩니다.

각 권장 예측자의 예측력은 목표가 연속형인지 또는 범주형인지에 따라 선형 회귀 또는 원시 Bayes 모형으로 계산됩니다.

필드 표

필드 처리 요약 주 보기에서 **목표**, **예측자** 또는 **예측자가 사용되지 않았음**을 클릭하면 필드 표 보기에 관련된 기능이 나열된 단순 표가 표시됩니다.

이 표에는 두 개의 열이 있습니다.

- **이름.** 예측자 이름.

  목표의 경우 목표가 변환되어도 필드의 원래 이름 또는 레이블이 사용됩니다.

  변환된 버전의 일반 예측자의 경우 설정 탭의 필드 이름 패널에서 이 이름은 사용자가 선택한 접두문자를 반영합니다(예: *_transformed*).

  날짜 및 시간에서 파생된 필드의 경우 최종 변환된 버전의 이름이 사용됩니다(예: *bdate_years*).

  구성된 예측자의 경우 구성된 예측자의 이름이 사용됩니다(예: *Predictor1*).
- **측정 수준.** 이는 데이터 유형을 나타내는 아이콘을 표시합니다.

  목표의 경우 **측정 수준**은 항상 변환된 버전을 반영합니다(목표가 변환된 경우). 예를 들어 순서(정렬된 세트)에서 연속(범위, 스케일)으로 변경하거나 그 반대로 변경합니다.

필드 세부사항

필드 주 보기에서 **이름**을 클릭하면 필드 세부사항 보기에 분포, 결측값 및 선택한 필드에 대한 예측력 도표가 표시됩니다. 또한 필드에 대한 처리 히스토리 및 변환 필드의 이름도 표시됩니다(적용 가능한 경우).

각 도표 세트의 경우 변환이 적용된 필드와 적용되지 않은 필드를 비교하도록 두 버전이 나란히 표시됩니다. 변환된 필드 버전이 없으면 도표가 원래 버전만 표시합니다. 파생된 날짜 또는 시간 필드와 구성된 예측자의 경우 도표는 새 예측자에 대해서만 표시됩니다.

참고: 범주가 너무 많아 필드가 제외된 경우 처리 히스토리만 표시됩니다.

분포 도표

연속형 필드 분포는 정규곡선으로 된 히스토리그램과 평균값에 대한 수직 참조선으로 표시됩니다. 범주형 필드는 막대도표로 표시됩니다.

히스토그램의 레이블은 표준편차 및 왜도를 표시하도록 지정됩니다. 그러나 값이 2이하이거나 원래 필드의 분산이 10-20보다 작은 경우 왜도는 표시되지 않습니다.

도표 위에 마우스를 올리면 히스토리그램의 평균 또는 범주에 대한 총 레코드 수와 퍼센트가 막대도표로 표시됩니다.

결측값 도표

원도표는 변환이 적용되거나 적용되지 않은 결측값의 퍼센트를 비교합니다. 도표 레이블은 퍼센트를 표시합니다.

ADP에서 결측값 처리를 실행한 경우 변환 후 도표에 대체값(결측값 대체에 사용된 값)이 레이블로 포함됩니다.

도표 위에 마우스를 올리면 결측값 수와 총 레코드수의 퍼센트가 표시됩니다.

예측력 도표

권장 필드의 경우 막대도표는 변환 전 후의 예측력을 표시합니다. 목표가 변환된 경우 예측력은 변환된 목표에 대하여 계산됩니다.

참고: 목표가 정의되지 않거나 주 보기 패널에서 목표를 클릭한 경우 예측력 도표가 표시되지 않습니다.

도표 위에 마우스를 올리면 예측력 값이 표시됩니다.

처리 히스토리 표

이 표는 변환된 필드 버전이 파생된 방법을 보여줍니다. ADP에서 수행한 동작은 실행된 순서대로 나열되지만 특정 단계의 경우 여러 동작이 특정 필드에 대해 실행되었을 수 있습니다.

참고: 이 표는 변환되지 않은 필드에 대해서는 표시되지 않습니다.

표에 있는 정보는 두 개 또는 세 개의 열로 나뉩니다.

- **동작.** 동작의 이름. (예: 연속 예측자). 자세한 정보는 『동작 세부사항』 주제를 참조하십시오.
- **세부사항.** 실행된 처리 목록. (예: 표준 단위로 변환).
- **함수.** 구성 예측자에 대해서만 표시된 경우 이는 입력 필드의 선형 조합을 표시합니다(예: 06*수명 + 1.21*높이).

동작 세부사항

동작 요약 주 보기에서 밑줄이 있는 **동작**을 선택하면 동작 세부사항 링크 보기에 실행된 각 처리 단계에 대한 동작 특정 및 공통 정보가 둘 다 표시됩니다. 동작 특정 세부사항이 먼저 표시됩니다.

각 동작의 경우 설명이 링크 보기의 위쪽에서 제목으로 사용됩니다. 동작 특정 세부사항은 제목 아래에 표시되고, 파생된 예측자 수, 필드 형변환 수, 목표 변환 수, 병합되거나 재정렬된 범주 수 및 구성되거나 제외된 예측장 수에 대한 세부사항이 포함될 수 있습니다.

각 동작이 처리될 때 처리에 사용된 예측자 수(예: 제외되거나 병합된 예측자 수)가 변경될 수 있습니다.

참고: 동작이 해제되거나 목표가 지정되지 않은 경우 동작 요약 주 보기에서 동작을 클릭하면 동작 세부사항 대신 오류 메시지가 표시됩니다.

9개의 가능한 동작이 있지만 모든 동작이 모든 분석에서 필수적으로 활성화되어 있지는 않습니다.

텍스트 필드 표

이 표는 다음 수를 표시합니다.

- 분석에서 제외된 예측자 수.

날짜 및 시간 예측자 표

이 표는 다음 수를 표시합니다.

- 날짜 및 시간 예측자에서 파생된 기간.
- 날짜 및 시간 요소 수.
- 파생된 총 날짜 및 시간 예측자 수.

날짜 기간이 계산되면 참조 날짜 또는 시간이 꼬리말로 표시됩니다.

예측자 숨기기 표

이 표는 다음 예측자를 처리에서 제외한 수를 표시합니다.

- 상수.
- 결측값이 너무 많은 예측자 수.
- 단일 범주에 너무 많은 케이스가 있는 예측자 수.
- 범주가 너무 많은 명목형 필드(세트) 수.
- 숨겨진 총 예측자 수.

측정 수준 확인 표

이 표는 필드 형변환 수를 표시하며 다음으로 나뉩니다.

- 연속형 필드로 형변환된 순서형 필드(정렬된 세트) 수.
- 순서형 필드로 형변환된 연속형 필드 수.
- 총 형변환 수.

연속형 또는 순서형의 입력 필드(목표 또는 예측자)가 없는 경우 꼬리말로 표시됩니다.

이상값 표

이 표는 이상값을 처리하는 방법에 대한 수를 표시합니다.

- 설정 탭에 있는 입력 및 목표 준비 패널의 사용자 설정에 따라 이상값을 찾거나 제거한 연속형 필드 수 또는 이상값을 찾고 결측으로 설정한 연속형 필드 수.
- 이상값 처리 후에 연속형 필드가 상수여서 제외된 수.

한 꼬리말은 이상값 분리점 값을 표시합니다. 반면 연속형의 입력 필드(목표 또는 예측값)가 없는 경우 다른 꼬리말이 표시됩니다.

결측값 표

이 표는 결측값이 대체된 필드 수를 표시하며 다음으로 나뉩니다.

- 목표. 이 행은 목표가 지정되지 않은 경우 표시되지 않습니다.
- 예측자. 이는 명목형(세트), 순서형(정렬된 세트) 및 연속형 수로 다시 나뉩니다.
- 대체된 총 결측값 수.

목표 표

이 표는 다음과 같이 목표가 변환되었는지 여부를 표시합니다.

- 정규성으로 Box-Cox 변환. 이는 특정 기준(평균과 표준편차)과 람다를 표시하는 열로 나뉩니다.
- 목표 범주가 재정렬되어 안정성을 향상시킵니다.

범주형 예측자 표

이 표는 범주형 예측자 수를 표시합니다.

- 안정성을 향상시키도록 가장 낮은 값에서 가장 높은 값으로 범주가 재정렬된 목표 수.
- 목표와의 연관성을 최대화하도록 범주가 병합된 목표 수.
- 희박한 범주를 처리하도록 범주가 병합된 목표 수.
- 목표와의 연관성이 낮아서 제외된 수.
- 병합 후에 상수여서 제외된 수.

범주형 예측자가 없는 경우 꼬리말이 표시됩니다.

연속 예측자 표

두 개의 표가 있습니다. 첫 번째는 다음 변환 수 중 하나를 표시합니다.

- 표준 단위로 변환된 예측자 값. 또한 이는 변환된 예측자 수, 지정된 평균 및 표준편차를 표시합니다.
- 공통 범위로 맵팽된 예측자 값. 또한 이는 지정된 최소 및 최대 값을 비롯하여 최소-최대 변환을 사용하여 변환된 예측자 수를 표시합니다.
- 구간화된 예측자 값과 구간화된 예측자 수.

두 번째 표는 예측자 수로 표시된 예측자 공간 구성 세부사항을 표시합니다.

- 구성됨.
- 목표와의 연관성이 낮아서 제외된 수.
- 구간화 후에 상수여서 제외된 수.
- 구성 후에 상수여서 제외된 수.

연속 예측자가 입력되지 않은 경우 꼬리말이 표시됩니다.

역변환 점수

ADP에서 목표를 변환한 경우 변환된 목표를 사용하여 작성된 이후의 모형은 변환된 단위로 점수가 지정됩니다. 결과를 해석하고 사용하려면 예측값을 다시 원래 척도로 변환해야 합니다.

1. 점수를 역변환하려면 메뉴에서 다음을 선택하십시오.

   변환 > 모형화를 위한 데이터 준비 > 역변환 점수...

2. 역변환할 필드를 선택하십시오. 이 필드에는 변환된 목표의 모형 예측값이 포함되어야 합니다.
3. 새 필드의 접미문자를 지정하십시오. 이 새 필드에는 변환하지 않은 목표의 원래 척도로 된 모형 예측값이 포함됩니다.
4. ADP 변환이 포함된 XML 파일 위치를 지정하십시오. 이는 대화형 또는 자동 데이터 준비 대화 상자에서 저장된 파일이어야 합니다. 자세한 정보는 17 페이지의 『변환 적용 및 저장』 주제를 참조하십시오.

제 5 장 특수 케이스 식별

비정상 탐지 프로시저는 해당 군집 집단의 노름(norm)에서의 편차를 기반으로 하는 특이 케이스를 검색합니다. 이 프로시저는 데이터 추정 분석 이전의 탐사 데이터 분석 단계에서 데이터 감사를 위해 특이 케이스를 신속하게 발견하도록 설계되었습니다. 이 알고리즘은 일반적인 비정상 탐지를 위해 설계되었으며, 이는 비정상 케이스 정의(예: 헬스케어 산업에서의 비정상적인 지불 패턴 탐지 또는 재무 산업에서의 돈세탁 탐지)가 비정상 정의가 잘 정의될 수 있는 특정 애플리케이션에 한정되지 않음을 의미합니다.

예제. 행정 처리 결과에 대한 예측 모형을 작성하기 위해 고용된 데이터 분석가는 해당 모형이 특이한 관측값에 민감할 수 있으므로 데이터 품질에 관심을 기울입니다. 이와 같이 범위를 벗어난 관측값 중 실제로 특별한 케이스를 나타내는 일부 값은 예측 모형을 작성하는 데 적합하지 않은 것으로 거를 수 있지만, 데이터 입력 오류로 인해 발생한 다른 관측값은 기술적으로 볼 때는 "올바른" 값이므로 데이터 검증 프로시저를 통해 발견할 수 없습니다. 특이 케이스 식별 프로시저는 분석가가 이를 처리하는 방법을 결정할 수 있도록 이러한 이상값을 찾아서 보고합니다.

통계. 이 프로시저는 동등 집단, 연속형 및 범주형 변수에 대한 동등 집단 기준, 동등 집단 기준의 편차를 기반으로 하는 비정상 지수, 특이한 케이스에 가장 기여하는 변수의 변수 영향 값을 생성합니다.

데이터 고려 사항

데이터. 이 프로시저는 연속형 변수와 범주형 변수에 모두 적용됩니다. 각 행은 고유한 관측값을 나타내고, 각 열은 동등 집단이 기반으로 하는 고유한 변수를 나타냅니다. 결과 표시를 위해 데이터 파일에서 케이스 식별 변수를 사용할 수 있지만 분석에서 사용할 수는 없습니다. 결측값은 허용됩니다. 지정한 가중변수는 무시됩니다.

탐지 모형은 새 검정 데이터 파일에 적용될 수 있습니다. 검정 데이터 요소는 학습 데이터 요소와 같아야 합니다. 또한 알고리즘 설정에 따라 모형설정에 사용된 결측값 처리은 점수화 이전에 검정 데이터 파일에 적용될 수 있습니다.

케이스 순서. 솔루션은 케이스 순서에 따라 다를 수 있습니다. 순서가 미치는 영향을 최소화하려면 케이스 순서를 무작위로 설정해야 합니다. 제공된 솔루션의 안정성을 확인하기 위해 다른 무작위 순서로 정렬된 케이스가 있는 여러 다른 솔루션을 확보하려고 할 수 있습니다. 파일이 너무 큰 솔루션에서는 다른 무작위 순서로 정렬된 샘플 케이스를 사용하여 실행을 여러 번 수행할 수 있습니다.

가정. 이 알고리즘은 모든 변수가 상수가 아니고 독립적이며, 케이스에 입력 변수에 대한 결측값이 없다고 가정합니다. 연속형 변수마다 개별 정규(가우시안)분포를 가지며 범주형 변수마다 다항분포 특성을 가진다고 가정합니다. 실제 내부 검정을 통해 이 프로시저가 독립성과 분산에 대한 가정에 그리 큰 영향을 받지 않는다는 결론을 얻었지만 이러한 가정을 충족하는 것이 좋습니다.

특이 케이스를 식별하려면 다음을 수행하십시오.

1. 메뉴에서 다음을 선택합니다.

   데이터 > 특이 케이스 식별...
2. 하나 이상의 분석 변수를 선택합니다.
3. 선택적으로 결과 레이블 지정에 사용할 케이스 식별자 변수를 선택합니다.

측정 수준을 알 수 없는 필드

측정 수준 경고는 데이터 세트에서 하나 이상의 변수(필드)에 대해 측정 수준을 알 수 없을 때 표시됩니다. 측정 수준은 이 프로시저의 계산 결과에 영향을 미치기 때문에 모든 변수에 정의된 측정 수준이 있어야 합니다.

데이터 스캔. 활성 데이터 세트의 데이터를 읽고 현재 알 수 없는 측정 수준이 있는 필드에 기본 측정 수준을 할당합니다. 데이터 세트가 큰 경우 시간이 걸릴 수 있습니다.

수동으로 할당. 알 수 없는 측정 수준이 있는 필드를 모두 나열하는 대화 상자를 엽니다. 이 대화 상자에서 해당 필드에 측정 수준을 할당할 수 있습니다. 데이터 편집기의 변수 보기에서도 측정 수준을 할당할 수 있습니다.

이 프로시저에 대해 측정 수준이 중요하기 때문에 모든 필드에 정의된 측정 수준이 있을 때까지는 대화 상자에 액세스하여 이 프로시저를 실행할 수 없습니다.

특이 케이스 결과 식별

특이 케이스 및 특이로 처리되는 원인 목록. 이 옵션은 세 개의 표를 생성합니다.

- 비정상 케이스 지수 목록은 특이로 식별되는 케이스를 표시하고 해당 비정상 지수 값을 표시합니다.
- 비정상 케이스 동등 ID 목록은 특이 케이스 및 해당 동등 집단에 관련된 정보를 표시합니다.
- 비정상 원인 목록은 케이스 번호, 원인변수, 변수 영향 값, 변수 값, 각 원인의 변수 노름(norm)을 표시합니다.

모든 표는 내림차순의 비정상 지수별로 정렬됩니다. 또한 변수 탭에 케이스 식별자 변수가 지정된 경우 케이스의 ID가 표시됩니다.

요약값. 이 그룹의 제어는 분포 요약을 생성합니다.

- **동등 집단 노름(norm).** 이 옵션은 연속형 변수 노름 표(분석에 연속형 변수가 사용된 경우)와 범주형 변수 노름 표(분석에 범주형 변수가 사용된 경우)를 표시합니다. 연속형 변수 노름 표는 각 동등 집단의 각 연속형 변수에 대한 평균과 표준편차를 표시합니다. 범주형 변수 노름 표는 각 동등 집단의 각 범주형 변수에 대한 최빈값(가장 널리 사용되는 범주), 빈도 및 빈도 퍼센트를 표시합니다. 연속형 변수의 평균 및 범주형 변수의 최빈값은 분석에서 노름 값으로 사용됩니다.
- **비정상 지수.** 비정상 지수 요약은 가장 특이한 항목으로 식별되는 케이스의 비정상 지수에 대한 기술통계량을 표시합니다.

- **분석 변수에 의한 원인 발생.** 이 표는 각 원인에 대해 각 변수의 원인 발생 빈도 및 빈도 퍼센트를 표시합니다. 또한 이 표는 각 변수의 영향에 대한 기술통계량을 보고합니다. 옵션 탭에서 최대 원인 수가 0으로 설정된 경우 이 옵션을 사용할 수 없습니다.
- **처리된 케이스.** 케이스 처리 요약은 활성 데이터 세트에 있는 모든 케이스, 분석에서 포함되거나 제외된 케이스, 각 동등 집단의 케이스에 대한 수 및 수 퍼센트를 표시합니다.

특이 케이스 식별 저장

저장할 변수. 이 그룹의 제어를 사용하여 모형 변수를 활성 데이터 세트에 저장할 수 있습니다. 또한 저장할 변수와 이름이 충돌하는 기존 변수를 대체하도록 선택할 수도 있습니다.

- **비정상 지수.** 각 케이스에 대한 비정상 지수 값을 지정된 이름의 변수에 저장합니다.
- **동등 집단.** 각 케이스의 동등 집단 ID, 케이스 빈도 및 *크기*(퍼센트)를 지정된 루트 이름의 변수에 저장합니다. 예를 들어 루트 이름이 *Peer*로 지정되면 *Peerid*, *PeerSize* 및 *PeerPctSize* 변수가 생성됩니다. *Peerid*는 동등 집단 ID 케이스이고, *PeerSize*는 집단의 *크기*이며, *PeerPctSize*는 집단의 *크기* 퍼센트입니다.
- **원인.** 지정된 루트 이름을 사용하여 원인 변수 세트를 저장합니다. 원인 변수 세트는 원인 변수 이름, 변수 영향 측도와 해당 값, 노름(norm) 값으로 구성됩니다. 세트 수는 옵션 탭에서 요청된 원인 수에 따라 다릅니다. 예를 들어, 루트 이름이 *Reason*으로 지정되면 *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* 및 *ReasonNorm_k* 변수가 생성됩니다. 여기서 *k*는 *k*번째 원인입니다. 원인 수가 0으로 설정되면 이 옵션을 사용할 수 없습니다.

모형 파일 내보내기. 모형을 XML 형식으로 저장할 수 있습니다.

특이 케이스 결측값 식별

결측값 탭에서 사용자 결측값과 시스템 결측값 처리를 제어할 수 있습니다.

- **분석에서 결측값 제외.** 결측값이 있는 케이스를 분석에서 제외합니다.
- **분석에 결측값 포함.** 연속형 변수의 결측값은 해당 총평균을 대체하고 범주형 변수의 결측 범주는 유효한 범주로 집단화되어 처리됩니다. 처리된 변수는 분석에 사용됩니다. 선택적으로 각 케이스에서 결측 변수의 비율을 나타내는 추가 변수를 작성하도록 요청하여 분석에서 해당 변수를 사용할 수 있습니다.

특이 케이스 식별 옵션

특이 케이스 식별 기준. 이러한 선택사항은 이상 목록에 포함된 케이스 수를 판별합니다.

- **비정상 지수 값이 가장 높은 케이스 퍼센트.** 100보다 작거나 같은 양의 정수를 지정하십시오.
- **비정상 지수 값이 가장 높은 케이스의 고정 숫자.** 분석에 사용되는 활성 데이터 세트에 있는 총 케이스 수보다 작거나 같은 양의 정수를 지정하십시오.
- **비정상 지수 값이 최소값을 충족하거나 초과하는 케이스만 식별.** 음수가 아닌 수를 지정하십시오. 비정상 지수 값이 지정된 분리점보다 *크거나* 같은 경우 이 케이스는 이상 항목으로 간주됩니다. 이 옵션은 **케이스 퍼**

센트 및 **케이스의 고정 숫자** 옵션과 함께 사용됩니다. 예를 들어 케이스의 고정 숫자를 50으로 지정하고 분리점 값을 2로 지정한 경우 비정상 목록은 비정상 지수 값이 각각 2보다 크거나 같은 최대 50개의 케이스로 구성됩니다.

동등 집단 수. 이 프로시저는 지정된 최소값과 최대값 사이의 동등 집단의 최적 수를 검색합니다. 이 값은 양의 정수여야 하고 최소값은 최대값을 넘지 않아야 합니다. 지정된 값이 같은 경우 프로시저는 동등 집단의 고정 숫자를 가정합니다.

참고: 데이터의 변동 크기에 따라 데이터에서 지원할 수 있는 동등 집단 수가 최소값으로 지정된 수보다 작은 상황이 발생할 수 있습니다. 이러한 경우 프로시저에서는 더 작은 동등 집단 수를 생성할 수 있습니다.

최대 원인 수. 원인은 변수 영향 측도, 이 원인에 대한 변수 이름, 변수 값 및 해당 동등 집단 값으로 구성됩니다. 음수가 아닌 정수를 지정하십시오. 이 값이 분석에 사용된 처리된 변수의 수와 같거나 큰 경우 모든 변수가 표시됩니다.

DETECTANOMALY 명령 추가 기능

명령 구문을 사용하여 수행할 수 있는 추가 기능은 다음과 같습니다.

- 모든 분석 변수를 명시적으로 지정하지 않고 분석에서 활성 데이터 세트에 있는 몇 가지 변수를 생략합니다(EXCEPT 부명령문 사용).
- 연속형과 범주형 변수 영향의 균형을 맞추도록 조정합니다(CRITERIA 부명령문에서 MLWEIGHT 키워드 사용).

명령 구문에 대한 자세한 내용은 *Command Syntax Reference*를 참조하십시오.

제 6 장 최적화 구간화

최적 구간화 프로시저는 각 변수의 값을 구간으로 분산시켜 하나 이상의 척도변수(이후부터 **입력 변수 구간화** 라고 함)를 이산화합니다. 구간 정보는 구간화 프로세스를 "관리하는" 범주형 안내변수에 관해 최적화됩니다. 구간은 추가 분석을 위해 원래 데이터 대신 사용할 수 있습니다.

예제. 변수에서 사용하는 구별된 값의 수를 줄이면 사용 수는 다음과 같습니다.

- 다른 프로시저의 데이터 요구 사항. 범주형 변수가 필요한 프로시저에서 사용하기 위해 이산화된 변수를 범주형으로 처리할 수 있습니다. 예를 들어 교차분석 프로시저에서는 모든 변수가 범주형이어야 합니다.
- 데이터 개인정보 보호정책. 실제 값 대신 구간화된 값을 보고하면 사용자 데이터 소스의 개인정보 보호정책을 보호하는 데 도움이 됩니다. 최적 구간화 프로시저는 구간 선택을 안내할 수 있습니다.
- 속도 성능. 일부 프로시저는 유일값이 적은 상태에서 작업할 때 더 효과적입니다. 예를 들어 이산화된 변수를 사용하여 다항 로지스틱 회귀 속도를 개선할 수 있습니다.
- 데이터의 완전한 또는 반완전한 분리 문제 발견.

최적 대비 시각적 구간화. 시각적 구간화 대화 상자에서는 안내 변수를 사용하지 않고 구간을 작성하는 몇 가지 자동화된 방법을 제공합니다. 이러한 "비지도" 규칙은 빈도표와 같은 기술통계량을 생성하는 데 유용하지만 최종 목적이 예측 모형 생성인 경우 최적 구간화가 우선합니다.

결과. 이 프로시저는 각 구간화 입력 변수의 구간 및 기술통계량에 대한 컷포인트 표를 생성합니다. 또한 구간화 입력 변수의 구간화된 값을 포함하여 활성 데이터 세트에 새 변수를 저장하고 구간화 규칙을 새 데이터 이산화에 사용하기 위한 명령 구문으로 저장할 수 있습니다.

최적 구간화 데이터 고려사항

데이터. 이 프로시저는 구간화 입력 변수가 척도 숫자변수일 수 있다고 예상합니다. 안내 변수는 범주형이어야 하고 문자열 또는 숫자일 수 있습니다.

최적 구간화를 확보하려면 다음을 수행하십시오.

1. 메뉴에서 다음을 선택합니다.

   변환 > 최적 구간화...
2. 하나 이상의 구간화 입력 변수를 선택합니다.
3. 안내 변수를 선택합니다.

기본적으로 구간화된 데이터 값이 포함된 변수는 생성되지 않습니다. 저장 탭을 사용하여 해당 변수를 저장하십시오.

최적 구간화 결과

결과 탭은 결과 표시를 제어합니다.

- **구간에 대한 끝점.** 각 구간화 입력 변수에 대한 끝점 세트를 표시합니다.
- **구간화된 변수에 대한 기술통계량.** 각 구간화 입력 변수에 대해 이 옵션은 유효한 값이 있는 케이스 수, 결측값이 있는 케이스 수, 고유한 유효값의 수, 최대값 및 최소값을 표시합니다. 안내 변수의 경우 이 옵션은 각 관련 구간화 입력 변수에 대한 계층 분포를 표시합니다.
- **구간화되는 변수에 대한 모형 엔트로피.** 각 구간화 입력 변수의 경우 이 옵션은 안내 변수와 관련된 변수의 예측 정확도에 대한 측도를 표시합니다.

최적 구간화 저장

변수를 활성 데이터 세트에 저장. 추가 분석에서는 원 변수를 사용하는 대신 구간화된 데이터 값을 포함하는 변수를 사용할 수 있습니다.

구간화 규칙을 구문으로 저장. 다른 데이터 세트를 구간화하는 데 사용할 수 있는 명령 구문을 생성합니다. 코딩변경 규칙은 구간화 알고리즘으로 식별되는 컷포인트를 기반으로 합니다.

최적 구간화 결측값

결측값 탭은 목록별 또는 대응별 삭제를 사용하여 결측값을 처리할지 여부를 지정합니다. 사용자 결측값은 항상 유효하지 않음으로 처리됩니다. 원 변수값을 새 변수로 코딩변경하면 사용자 결측값이 시스템 결측값으로 변환됩니다.

- **대응별.** 이 옵션은 각 안내 및 구간화 입력 변수 쌍에서 수행됩니다. 이 프로시저는 안내 및 구간화 입력 변수에서 결측되지 않은 값이 있는 모든 케이스를 사용합니다.
- **목록별** 이 옵션은 변수 탭에 지정된 모든 변수에서 수행됩니다. 케이스에서 변수가 누락된 경우 전체 케이스가 제외됩니다.

최적 구간화 옵션

전처리. 유일값이 많은 구간화 입력 변수를 "사전 구간화"하면 최종 구간의 품질에 영향을 미치지 않고 처리 시간을 향상시킬 수 있습니다. 최대 구간 수는 작성된 구간 수에 대한 상한입니다. 따라서 최대값을 100으로 지정했지만 구간화 입력 변수가 유일값 1000개보다 작은 경우 구간화 입력 변수에 대해 작성된 전처리된 구간 수가 구간화 입력 변수의 유일값 수와 같습니다.

희박하게 채워진 구간. 때때로 프로시저에서 아주 적은 케이스가 있는 구간이 생성될 수 있습니다. 다음 처리 방법은 이러한 pseudo 컷포인트를 삭제합니다.

주어진 변수에 대해 알고리즘이 n개의 $_{최종}$ 컷포인트와 n개의 $_{최종}$+1 구간을 찾았다고 하십시오. i = 2, ..., n개의 $_{최종}$ 구간(두 번째 최상위 값 구간을 통해 지정된 두 번째 최하위 값 구간)의 경우 다음을 계산하십시오.

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

여기서 *sizeof(b)*는 구간의 케이스 수입니다.

이 값이 지정된 병합 임계값보다 작은 경우 b_i는 희박하게 채워진 값으로 간주되고 b_{i-1} 또는 b_{i+1}과 병합됩니다(하한 클래스 정보 엔트로피가 있음).

이 프로시저는 구간을 통해 단일 전달을 작성합니다.

구간 끝점. 이 옵션은 간격의 하한 정의 방법을 지정합니다. 이 프로시저는 컷포인트 값을 자동으로 판별하므로 이는 선호의 문제입니다.

첫 번째(가장 낮은)/마지막(가장 높은) 구간. 이러한 옵션은 각 구간화 입력 변수에 대한 최소 및 최대 컷포인트 정의 방법을 지정합니다. 일반적으로 이 프로시저에서는 구간화 입력 변수에 실수의 값을 사용할 수 있지만 범위 제한에 대한 일부 이론 또는 실행 상의 이유가 있는 경우 최소값/최대값으로 한계를 지정할 수 있습니다.

OPTIMAL BINNING 명령 추가 기능

명령 구문을 사용하여 수행할 수 있는 추가 기능은 다음과 같습니다.

- 동일한 빈도 방법을 통해 비지도 구간화를 수행하십시오(CRITERIA 부명령문 사용).

명령 구문에 대한 자세한 내용은 *Command Syntax Reference*를 참조하십시오.

주의사항

이 정보는 미국에서 제공되는 제품 및 서비스용으로 작성된 것입니다. 본 자료는 다른 언어로도 제공될 수 있습니다. 그러나 자료에 접근하기 위해서는 해당 언어로 된 제품 또는 제품 버전의 사본이 필요할 수 있습니다.

IBM은 다른 국가에서 이 책에 기술된 제품, 서비스 또는 기능을 제공하지 않을 수도 있습니다. 현재 사용할 수 있는 제품 및 서비스에 대한 정보는 한국 IBM 담당자에게 문의하십시오. 이 책에서 IBM 제품, 프로그램 또는 서비스를 언급했다고 해서 해당 IBM 제품, 프로그램 또는 서비스만을 사용할 수 있다는 것을 의미하지는 않습니다. IBM의 지적 재산권을 침해하지 않는 한, 기능상으로 동등한 제품, 프로그램 또는 서비스를 대신 사용할 수도 있습니다. 그러나 비IBM 제품, 프로그램 또는 서비스의 운영에 대한 평가 및 검증은 사용자의 책임입니다.

IBM은 이 책에서 다루고 있는 특정 내용에 대해 특허를 보유하고 있거나 현재 특허 출원 중일 수 있습니다. 이 책을 제공한다고 해서 특허에 대한 라이센스까지 부여하는 것은 아닙니다. 라이센스에 대한 의문사항은 다음으로 문의하십시오.

150-945
서울특별시 영등포구 국제금융로 10, 3IFC
한국 아이.비.엠 주식회사
대표전화서비스: 02-3781-7114

2바이트(DBCS) 정보에 관한 라이센스 문의는 한국 IBM에 문의하거나 다음 주소로 서면 문의하시기 바랍니다.

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM은 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여(단, 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 이 책을 "현상태대로" 제공합니다. 일부 지역에서는 특정 상거래에 있어 명시적 또는 묵시적 보증 책임에 대한 불인정을 허용하지 않으므로, 이런 지역에서는 위의 사항이 적용되지 않습니다.

이 정보에는 기술적으로 부정확한 내용이나 인쇄상의 오류가 있을 수 있습니다. 이 정보는 주기적으로 변경되며, 변경된 사항은 최신판에 통합됩니다. IBM은 이 책에서 설명한 제품 및/또는 프로그램을 사전 통지 없이 언제든지 개선 및/또는 변경할 수 있습니다.

이 정보에서 언급되는 비IBM의 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 해당 웹 사이트의 자료는 본 IBM 제품 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

IBM은 귀하의 권리를 침해하지 않는 범위 내에서 적절하다고 생각하는 방식으로 귀하가 제공한 정보를 사용하거나 배포할 수 있습니다.

(i) 독립적으로 작성된 프로그램과 기타 프로그램(본 프로그램 포함) 간의 정보 교환 및 (ii) 교환된 정보의 상호 이용을 목적으로 본 프로그램에 관한 정보를 얻고자 하는 라이센스 사용자는 다음 주소로 문의하십시오.

150-945
서울특별시 영등포구 국제금융로 10, 3IFC
한국 아이.비.엠 주식회사
대표전화서비스: 02-3781-7114

이러한 정보는 해당 조건(예를 들면, 사용료 지불 등)하에서 사용될 수 있습니다.

이 정보에 기술된 라이센스가 부여된 프로그램 및 프로그램에 대해 사용 가능한 모든 라이센스가 부여된 자료는 IBM이 IBM 기본 계약, IBM 프로그램 라이센스 계약(IPLA) 또는 이와 동등한 계약에 따라 제공한 것입니다.

인용된 성능 데이터와 고객 예제는 예시 용도로만 제공됩니다. 실제 성능 결과는 특정 구성과 운영 조건에 따라 다를 수 있습니다.

비IBM 제품에 관한 정보는 해당 제품의 공급업체, 공개자료 또는 기타 범용 소스로부터 얻은 것입니다. IBM에서는 이러한 제품들을 테스트하지 않았으므로, 비IBM 제품과 관련된 성능의 정확성, 호환성 또는 기타 청구에 대해서는 확신할 수 없습니다. 비IBM 제품의 성능에 대한 의문사항은 해당 제품의 공급업체에 문의하십시오.

IBM이 제시하는 방향 또는 의도에 관한 모든 언급은 특별한 통지 없이 변경될 수 있습니다.

이 정보에는 일상의 비즈니스 운영에서 사용되는 자료 및 보고서에 대한 예제가 들어 있습니다. 이들 예제에는 개념을 가능한 완벽하게 설명하기 위하여 개인, 회사, 상표 및 제품의 이름이 사용될 수 있습니다. 이들 이름은 모두 가공의 것이며 실제 기업 및 인물과 유사하더라도 이는 전적으로 우연입니다.

저작권 라이센스:

이 정보에는 여러 운영 플랫폼에서의 프로그래밍 기법을 보여주는 원어로 된 샘플 응용프로그램이 들어 있습니다. 귀하는 이러한 샘플 프로그램의 작성 기준이 된 운영 플랫폼의 애플리케이션 프로그래밍 인터페이스(API)에 부합하는 애플리케이션을 개발, 사용, 판매 또는 배포할 목적으로 IBM에 추가 비용을 지불하지 않고 이들 샘플 프로그램을 어떠한 형태로든 복사, 수정 및 배포할 수 있습니다. 이러한 샘플 프로그램은 모든 조건하에서 완전히 테스트된 것은 아닙니다. 따라서 IBM은 이러한 프로그램의 신뢰성, 서비스 가능성 또는 기능을 보증하거나 진술하지 않습니다. 본 샘플 프로그램은 일체의 보증 없이 "현상태대로" 제공됩니다. IBM은 귀하의 샘플 프로그램 사용과 관련되는 손해에 대해 책임을 지지 않습니다

이러한 샘플 프로그램 또는 파생 제품의 각 사본이나 그 일부에는 반드시 다음과 같은 저작권 표시가 포함되어야 합니다.

© (귀하의 회사명) (연도). 이 코드의 일부는 IBM Corp.의 샘플 프로그램에서 파생됩니다.

상표

IBM, IBM 로고 및 ibm.com은 전세계 여러 국가에 등록된 International Business Machines Corp.의 상표 또는 등록상표입니다. 기타 제품 및 서비스 이름은 IBM 또는 타사의 상표입니다. 현재 IBM 상표 목록은 "저작권 및 상표 정보" 웹 페이지(www.ibm.com/legal/copytrade.shtml)에 있습니다.

Adobe, Adobe 로고, PostScript 및 PostScript 로고는 미국 및/또는 기타 국가에서 사용되는 Adobe Systems Incorporated의 등록상표 또는 상표입니다.

Intel, Intel 로고, Intel Inside, Intel Inside 로고, Intel Centrino, Intel Centrino 로고, Celeron, Intel Xeon, Intel SpeedStep, Itanium 및 Pentium은 미국 또는 기타 국가에서 사용되는 Intel Corporation 또는 그 계열사의 상표 또는 등록상표입니다.

Linux는 미국 또는 기타 국가에서 사용되는 Linus Torvalds의 등록상표입니다.

Microsoft, Windows, Windows NT 및 Windows 로고는 미국 또는 기타 국가에서 사용되는 Microsoft Corporation의 상표입니다.

UNIX는 미국 및 기타 국가에서 사용되는 The Open Group의 등록상표입니다.

Java 및 모든 Java 기반 상표와 로고는 Oracle 및/또는 그 계열사의 상표 또는 등록상표입니다.

색인

[가]

감독 구간화
  대비 비지도 구간화 31
  최적 구간화 31
검증 규칙 3
검증 규칙 위반
  데이터 검증 10
검증 규칙 정의 3
  교차-변수 규칙 4
  단일-변수 규칙 3
결측값
  특이 케이스 식별 29
교차-변수 검증 규칙
  검증 규칙 정의 4
  데이터 검증 9
구간에 대한 끝점
  최적 구간화 32
구간화 규칙
  최적 구간화 32
기간 계산
  자동 데이터 준비 13

[다]

단일-변수 검증 규칙
  검증 규칙 정의 3
  데이터 검증 9
대화형 데이터 준비 11
데이터 검증 7
  결과 9
  교차-변수 규칙 9
  기본 확인 8
  단일-변수 규칙 9
  데이터 검증 7
  저장할 변수 10
동등 집단
  특이 케이스 식별 28, 29

[마]

모형 보기
  자동 데이터 준비 17

[바]

변수 구성
  자동 데이터 준비 16
분석 가중치
  자동 데이터 준비 15
불완전한 케이스 식별자
  데이터 검증 10
비어 있는 케이스
  데이터 검증 10
비정상 지수
  특이 케이스 식별 28, 29
비지도 구간화
  대비 감독 구간화 31

[사]

사전 구간화
  최적 구간화 32
순환 시간 요소
  자동 데이터 준비 13

[아]

연속형 목표 표준화 15
원인
  특이 케이스 식별 28, 29

[자]

자동 데이터 준비 11
  날짜 및 시간 준비 13
  데이터 품질 향상 14
  동작 세부사항 22
  동작 요약 20
  모형 보기 17
  목적 11
  변수 구성 16
  변환 적용 17
  보기 간 링크 17
  보기 재설정 17
  연속형 목표 표준화 15
  예측력 20
  이름 필드 16
자동 데이터 준비 *(계속)*
  점수 역변환 25
  측정 수준 수정 14
  필드 12
  필드 변환 15
  필드 분석 19
  필드 세부사항 21
  필드 제외 14
  필드 조정 15
  필드 처리 요약 18
  필드 표 21
  필드선택 16
중복 케이스 식별자
  데이터 검증 10

[차]

최적화 구간화 31
  결과 32
  결측값 32
  옵션 32
  저장 32

[타]

특수 케이스 식별 27
  결과 28
  결측값 29
  모형 파일 내보내기 29
  옵션 29
  저장할 변수 29

[파]

필드선택
  자동 데이터 준비 16

B

Box-Cox 변환
  자동 데이터 준비 15

M

MDLP
최적 구간화 31